19th ICCRTS

C2 Agility: Lessons Learned from Research and Operations

Providing Agility in C2 Environments Through Networked Information Processing: A Model of Expertise

Topic(s)

Topic 4: Experimentation, Metrics, and Analysis

Name of Author(s)

Kevin Chan, Jin-Hee Cho
Network Science Division, Computational & Information Sciences Directorate
US Army Research Laboratory
Adelphi, MD

Sibel Adali
Rensselaer Polytechnic Institute
Department of Computer Science
Troy, NY

Jennifer Mangels, Olta Hoxha, Ksenia Lila
Baruch College
City University of New York
New York, NY

Damon Abraham
Doctoral Program in Psychology
Graduate Center of the City of New York
New York, NY

Point of Contact

Kevin Chan
kevin.s.chan.civ@mail.mil
+1 301 394 5640

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Providing Agility in C2 Environments Through Networked Information Processing: A Model of Expertise	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **US Army Research Laboratory,Network Science Division, Computational & Information Sciences Directorate,Adelphi,MD,20783**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the 18th International Command & Control Research & Technology Symposium (ICCRTS) held 16-19 June, 2014 in Alexandria, VA.

14. ABSTRACT
The amount of information and the diversity of sources of information in tactical environments continue to grow, increasing the requirement of intelligence analysts to filter the valuable information from the noise. Information is often gathered by individuals of differing expertise in a given topic area. Furthermore, it might be inherently more difficult to determine signal from noise in some topics. Hence, it would be crucial to understand how both the level of expertise and the level of difficulty of a problem impact the ability of a person to correctly classify it. By understanding the distribution of errors, it is possible to create methods to overcome them. To accomplish this, we analyze the results of an experimental study and develop a mathematical model for expertise. To the best of our knowledge, no such model exists. Based on this model, we show how we can develop an agent simulation that mimics the expected performance of a human agent under different conditions.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**35**	

Abstract

The amount of information and the diversity of sources of information in tactical environments continue to grow, increasing the requirement of intelligence analysts to filter the valuable information from the noise. Information is often gathered by individuals of differing expertise in a given topic area. Furthermore, it might be inherently more difficult to determine signal from noise in some topics. Hence, it would be crucial to understand how both the level of expertise and the level of difficulty of a problem impact the ability of a person to correctly classify it. By understanding the distribution of errors, it is possible to create methods to overcome them. To accomplish this, we analyze the results of an experimental study and develop a mathematical model for expertise. To the best of our knowledge, no such model exists. Based on this model, we show how we can develop an agent simulation that mimics the expected performance of a human agent under different conditions.

1. Motivation

Decision making in command and control (C2) environments requires gathering, processing and sharing of all available information to establish intelligence and to increase situation awareness. Given the amount of limited time and vast amounts of information and intelligence, it is necessary to understand how accurately one may make decisions on information given particular circumstances. Often complex criteria play a role in such a decision: trustworthiness (reliability, benevolence) of the sources, ability (expertise) of analysts to analyze and filter the information, the underlying credibility of the findings, and other network level factors like social influence. In this paper, we study the expertise component of this process. We develop a mathematical model of expertise based on an experimental study and discuss how this model can be incorporated into an agent model for information processing in networks.

Expertise is one of the main determinants of how well information will be judged as correct or incorrect by an analyst. In a networked decision making scenario, this results in the information to be disseminated or filtered out at individual nodes. Any errors that are propagated within the network will lead to both extra work and also to possibly false conclusions. Hence, it is important to understand how different factors impact accuracy in decisions at differing levels of expertise. There are various models of expertise, but these models tend to be qualitative: they describe in which way an expert's decision making differs from a person with no expertise in a topic. However, such models do not provide a good insight into how expertise can be modeled mathematically in an agent simulation. How can we model the errors a person makes based on their level of expertise? What mathematical model best approximates the type of errors expected?

To solve this problem, we choose a data driven approach. We analyze results from a set of ongoing experimental studies on information processing, and various results of these experiments are studied to answer these questions. In particular, we consider three main moderating factors: difficulty of a given problem domain, the amount of information available to make a decision and time pressure. We also consider the impact of self-reported domain expertise in these experiments. We look at experiments that vary these factors to understand the impact they have on the accuracy of decisions. The results presented here are an initial analysis of the experimental results and a development of the model for expertise and information sharing and processing in C2 environments. We first describe the military relevance and connections to intelligence gathering of this study. We then describe the experimental paradigm and some of the initial insights obtained from earlier experiments conducted using this paradigm. The next section discusses an initial analysis of the experiments and development of a model for agents in intelligence scenarios.

1.1. Information Sharing Scenario

The increased variation and volume of sources in intelligence operations is resulting in the increased complexity and difficulty to process information correctly and efficiently. This has been an area of active research, particularly within the C2 research community. One example of an experimental approach is ELICIT (Experimental Laboratory for Investigating Collaboration,

Information-sharing, and Trust), a platform designed to study organizational behavior and performance [1]. Further, this work has resulted in the development of software agent models to perform similarly to human participants in ELICIT experiments. Currently, the software agents indiscriminately accept the accuracy and value of the factoids coming from cooperating agents in the experiments. In operational situations, there is undoubtedly variation in the expertise and capability of information sources. Understanding the behavior and tendencies of both human and automated sources of information and intelligence to make mistakes will allow both analysts and commanders to improve their decision making performance. Such information sharing scenarios are present in tactical situations. Two commonly referenced networked environments are COIST (Company Intelligence Support Team) and coalition networks. In both of these situations, there is a large variation in the trust and capability of entities with which commanders must interact to accomplish their mission. The experiments run will provide insight into how well individuals are able to process information based on the previously mentioned experimental parameters. This will affect the rate of information flow

1.2 Background and Related Work

In the research domains of *computer science* and *artificial intelligence*, expert systems or expert locator systems have been developed to maximize efficiency and effectiveness for the use of expertise in networks. Yagi et al. [4] propose a decision-making system based on expertise knowledge in orthodontics. Meyer and Booker [5] discuss the techniques for eliciting and analyzing expert judgment for various unique situations and work areas. In addition, Booker and Meyer [6] introduce probabilistic and fuzzy logic based theories that have been adopted to make decisions under various uncertain situations where information and source credibility is critical. Chu-Carroll and Carberry [8] present a computational framework that initiates expert-consultation sub-dialogues to solve a problem on whether or not an agent accepts or rejects a proposal by the other agent.

In addition, Huang et al. [11] propose expertise management systems in organizations to support expertise information collection, processing, and distribution by developing a visualization technique exploring the expertise space. Li et al. [14] identify expertise networks in online communities based on textual information describing users' feedback and social connections. For designing customized expert locator systems, Nevo et al. [15] examine different experts' attributes in two different contexts of expertise seeking: knowledge allocation and knowledge retrieval. When expert seekers retrieve knowledge from the knowledge retrieval, they place higher emphasis on the expertise level of a source. On the other hand, when expertise seekers choose an expert to transfer knowledge to the knowledge allocation, they pay more attention to the expert's benevolence. As seen in [4], [5], [6], [8], [10], [12], [13], [14], [15], although it is proposed that diverse intelligent expertise locator systems find the correct expertise in particular domains, little work has been done to examine the relationship between the expertise of an agent and its decision making ability.

In *information systems research*, information-sharing behaviors have been studied. According to Constant et al. [7], in work settings, information can be perceived as product (like a commodity; often called 'tangible information' if it is a written document) or expertise (called 'intangible information' if it is an unwritten information based on individual experience/background). In

work environments, information sharing behavior can be based on reciprocal relationships. However, expertise can mean more than a simple commodity and imply a person's 'identity' and 'self-worth.' Sharing expertise provides pragmatic benefits as well as the expression and consistency of the individual's identity and value. In this sense, sharing expertise can provide personal benefits promoting self-esteem, pride, self-efficacy, personal identification with colleagues and organizations, obtaining a better reputation and increasing commitment. The authors' experiments support the idea that sharing tangible information in work settings may be affected by prosocial attitudes and norms of organizational ownership while sharing expertise is more related to people's own self-expressive needs.

In *management/marketing research*, relationship between expertise, source credibility, and information sharing behaviors have been studied. Thomas-Hunt et al. [9] examine how social connectedness and perceived expertise affect the emphasis of unique and shared knowledge within functionally heterogeneous groups consisting of experts and non-experts. When an expert is less socially connected, he/she is more likely to share unique knowledge. On the other hand, socially connected experts tend to emphasize shared knowledge and other members' unique knowledge contributions. Harmon and Coney [23] look at how source credibility can affect persuasiveness in buy and lease situations. They show that the impact of source credibility on attitude and behavioral intention depends on the situation. In marketing research, the correlation between trustworthy source and information credibility is commonly assumed. However, Wiener and Mowen [21] show in their experiments that expertise has more influence on information credibility than trustworthiness of sources as expert sources affect perceptions of the product's qualities. Similarly, the significant relationship between source credibility and expertise / attractiveness is also studied by McGinnes and Ward [22]. Although [9], [21], [22] provides good insight on the relationship between expertise and social connectedness or information credibility, there has been little effort that investigates how an entity's expertise affects decision making performance.

In *organizational behavior* and *human decision making research*, interesting findings on relations between trust, expertise, confidence, and accuracy have been explored. Sniezek and Buckley [18] study the impact of advice on a judge's own initial choice. The experiments are set up with teams consisting of one judge and two advisors. They are given by a task exposed of 70 items with two alternatives each, to make a final choice. Judges make final team choices and provide confidence assessments under one of the following conditions: dependent, cued, and independent. *Dependent* is a condition that a judge has no basis for choice; *cued* is that a judge selects only after an advice is given; and *independent* means that a judge makes his/her own tentative choice before an advice is given as well as subsequent final choice. Their findings show that *independent* judges performed the best and *dependent* judges performed the worst. Conflict of advisors' opinions generated less confidence in the advice, which affected the judge's final choice. Consensus of advisors' opinions increased the judge's confidence in the advice and increasingly affected the judge's choice. In Sniezek et al. [16], judges make a final decision based on all information provided by advisors. If an advisor has high confidence in his/her expertise, a judge has high trust in the advisor and is more likely to take the advice from the advisor. Yaniv [17] investigates the impact of advice on judgment and the consequences of the use of the advice for judgment accuracy. In the experiments, respondents are asked to give final judgments based on their initial opinions and an advice is presented to them. The results show

that (1) the respondents weighted their own opinion more than the advice; (2) the respondents with more expertise discounted the advice more; (3) the respondents weighted the advice that is more distant from their initial opinion less; and (4) using the advice has improved accuracy significantly but not optimally. This work is similar with our work in that a user can use an alternative answer to make a final decision. However, in our experiments, in addition to providing one alternative answer, three alternative answers are also given as another condition. Besides, we vary the difficulty of a problem and time pressure under these two cases of providing alternative answers.

In *psychology*, Perfect et al. [20] study the relation between confidence and accuracy for general knowledge and eyewitness memory. Their findings show that there exists correlation between confidence and accuracy for general knowledge but not for eyewitness memory. Similarly, Sporer et al. [19] support the weak relationship between confidence and eyewitness memory in their study. They conduct a meta-analytic review of 30 studies based on staged-event methods with "target-present" and "target-absent" lineups. The results show that when choosers make positive identification, the correlation between confidence and accuracy was consistently high. Besides correct choosers have a higher mean confidence level than incorrect choosers for all studies. Similar to [19], [20], our work also looks at the relation between confidence (as a measure of expertise) and accuracy. But our work uses two different ways of capturing 'confidence' as a measure of expertise in terms of self-reported expertise and recall expertise (i.e., average accuracy in the initial recall phase of the experiment) in order to enhance validity of measuring 'confidence' related to actual expertise.

2. Experimental Paradigm

In this paper, we consider an experimental platform to study the expertise of individuals within a question answering setting involving general knowledge domain questions. While this situation does not replicate the operational situations, this focuses on the individual and its ability to accurately solve problems under certain situations. The experimental paradigm that we study here is developed by CUNY (City University of New York) Baruch College over many years and has been used in many different studies. The experiments make use of a set of general knowledge questions, 433 in total. Each question is carefully phrased so that there is only a single answer and the answer consists of a single word. The questions have differing difficulty. The general difficulty of the questions has been established using a normative study in which the questions were asked to a group of 283 students from Baruch College. The question difficulty (d_j) metric obtained from this study corresponds to the percentage of subjects who answered the question correctly. Higher values mean higher difficulty ($d_j = 1$ means 0% of the participants answered the question correctly, while $d_j = 0$ means 100% of the participants answered the question correctly). As this normative data is obtained from the Baruch College students, all the subsequent experiments are also conducted using the same subject pool.

The experiment is setup such that each participant is asked a series of questions. For each question, the subject is asked to give an initial answer. This initial answer corresponds to a cued recall test. The only retrieval cue the subjects have at this stage is the question. After the subject answers the question, they are asked for their confidence in their answer. Next, the participant is

given either one or three alternate answers from "other players" who have performed the task previously. They are allowed to keep their initial answer or choose one of the alternate answers as their final answer. The subject is asked for their confidence in their final answer. Finally, the subject is shown the correct answer. If their final answer is correct, the subject is notified that it has received points for a correct answer. An incorrect answer is not awarded any points. Then, the next question is asked (Figure 1).

The choice of alternative answers comes from a database of answers collected using a set of experiments prior to the ones we report on here. These early experiments were highly similar to the present experiment in their initial recall phase. The database contains the relative frequency with which specific answers, both the correct and incorrect answers, are given to a specific question. Because questions in the database were not asked equally often, answer frequency was normalized within each question. The database was manually cleaned to exclude answers that were truly noise, such as "I don't know," slang and repeats of words in the question that obviously were not possible answers to the question.

We are especially interested in the distribution of incorrect answers. In fact, there are quite a lot of questions in our database for which the most frequent answer is not the correct answer. Hence, one can argue that a source of difficulty for questions is that they have a very strong lure (i.e., a very plausible answer). Overall, high frequency wrong answers tend to be strongly related to the correct answer. They may also be high frequency words in the lexicon, and therefore, have greater retrieval fluency in general. Another source of difficulty for questions is the number of possible answers. For example, a question asking for a continent has only few possible answers that are easy to enumerate. However, a question asking for a proper first name may have many possible answers. Such a question may be inherently difficult unless one knows the correct answer. Quantifying problem difficulty is an ongoing research problem. In this paper, we will concentrate on an overall quantitative measure of difficulty as the proportion of times the correct answer was given and investigate other qualitative factors leading to this difficulty in future work.

In the experiments that we report on here, the questions are organized into three main categories:

- Math & Sciences
- Arts & Humanities
- History & Geography

These categories are tuned carefully so that the various statistics for each category is very similar. These include the mean and standard deviation of difficulty, and the total number of questions within the category. Before any questions are asked, the experiment subjects are first asked to rank their own expertise in the three topics: from the topic they feel most knowledgeable into the topic they feel the least expert in. The subjects judge their own expertise by ordering these three topics. Three most frequent incorrect answers are chosen for each question as possible alternate answers. These answers are associated with a level of credibility based on the frequency with which that answer had been given in the answer database.

The participants receive raffle tickets based on the total number of correct answers that they receive. We vary two different factors. Experiments can be "timed" or not. In the "timed" version of the study, participants lose earnings the longer it takes them to choose their final answer (timer starts when alternative answers are first presented). Hence, there is a level of time pressure that potentially impacts how they choose their final answer. The task-wide accuracy of the initial answer is titrated to 33% correct, by varying the difficulty of the questions presented. The titration in these experiments has been very successful. Hence, the overall performance of the subjects remains roughly similar for initial answers, but their underlying difficulty changes.

The second factor we vary is whether one or three alternate answers are shown after the initial recall phase. In the one alternative answer scenario, the alternative following an incorrect initial answer is correct 50% of the time, such that the maximum final success rate is 66%. One third of the time both the participant and the alternative are incorrect. In the three alternate answer scenario, one of the four answers on the screen (including the three alternatives and the participant's own answer) is always correct, so it is possible in this scenario for the final success rate to reach 100%. Thus, in order to make final accuracy data for the one and three alternative versions of the task comparable, we will focus our analysis on trials in which the correct answer was an option (66% of one alternative trials, 100% of three alternative trials).

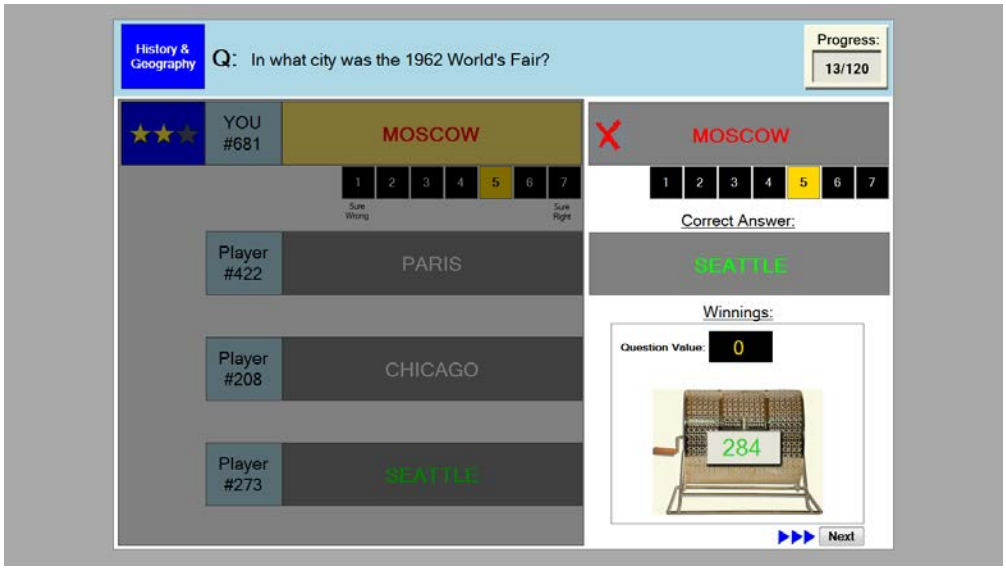


Figure 1. Screen shot of Trivia Experiment seen by the subject.

Figure 1 is a screen capture of the application, where the question is "In what city was the 1962 World's Fair?" with the initial response of "MOSCOW". The initial confidence of this subject was 5 out of 7. The subject is given three alternate answers of "PARIS", "CHICAGO", and "SEATTLE". The final answer chosen was "MOSCOW" (subject chose not to switch from their initial answer) and the confidence remained at 5. However, the correct answer was "SEATTLE". Note that the initial answer and alternative are greyed out during the presentation of the correct answer and winnings outcome in order to focus attention on that part of the screen. These sections are lit brightly during the initial answer and final decision phases.

In this paper, we present results from all four different versions of the experiment discussed. The experiments were run at Baruch College in 2013. Each experiment is conducted for as long as the participants need to answer all 120 questions and the total number of subjects tested for each treatment is in Table 1. Also note that different subjects were used for each of the variations.

Test	Treatment	Abbreviation	# of subjects
(1-a)	One alternate answer with no time pressure	*alt1off*	33
(1-b)	One alternate answer with time pressure	*alt1on*	39
(3-a)	Three alternate answers with no time pressure	*alt3off*	39
(3-b)	Three alternate answers with time pressure	*alt3on*	37

Table 1: Description of experiment treatments and number of subjects tested.

2.2. Definition of Expertise

We have a number of different possible definitions of expertise based on this experimental data.

- **Self reported expertise**: a scale of 1-3 depending on whether the subject chose the topic as his or her top, next lowest choice. Score 1 corresponds to the highest level of expertise.
- **Recall expertise**: Average difficulty of questions answered correctly in the initial recall phase of the experiment.
- **Meta-cognitive recall expertise**: To which degree the reported level of confidence for the initial response correlates with the probability that the answer is correct.
- **Selection expertise**: The accuracy of the final answer in the three answer version of the experiment. In this case, we have chosen the test cases in which they are guaranteed to have seen the correct answer. Hence, their expertise corresponds to their ability to pick the correct answer from a set of answers, including their own.
- **Meta-cognitive selection expertise**: The correlation between the accuracy in the final answer and the reported confidence, in the three answer version as well.

Note that each measure has certain drawbacks. For example, a number of measures do not take into account the difficulty of the question. Similarly, the self reported expertise lacks precision. A person may have two topics of expertise, but they are forced to use only one in this setting. Also our setting is limited in modeling expertise in a few dimensions. The subject pool is not guaranteed to have individuals who are truly experts in a given topic. Furthermore, the setting of trivia questions limits the type of expertise considered. For example, this setting is not appropriate for modeling expertise in analytical problem solving that would be required in some settings. An example of such a setting would be ranking different options by considering different pros and cons. However, it allows us to consider cases where individuals have to judge information as fact or noise with varying levels of familiarity with the underlying problem domain.

In the remainder of this paper, we will focus on the self-reported expertise and recall expertise. The study of meta-cognition expertise related to confidence is a topic of future study.

3. Analysis of Experimental Results

In this section, we study the relationship between expertise, problem difficulty, time pressure, the amount of information available and accuracy of answers. We will consider different definitions of expertise. We then propose a mathematical model of expertise that builds on our findings.

In terms of filtering of the results, we consider question difficulty d_j in bins of 0.2 increments. The questions with the highest difficulty have difficulty between 0.8 and 1.0. In particular, we would like to understand how question difficulty impacts accuracy of decisions, given the other factors. Note that we will make no distinction between questions with strong lures and questions with a large range of responses. This is a topic we would like to study in future work.

3.1 Accuracy and Problem Difficulty

First, we look at the accuracy of the responses as a function of problem difficulty. We are to expect a negative correlation between accuracy and problem difficulty. In looking at the initial accuracy, we would also expect that the results are identical for each of the treatments, as at this stage the experiments are identical. Figure 2 (a) is a plot of this relationship, which shows the negative correlation and identical relationships. Obviously the time pressure and alternate answers are not a factor at this point of the experiment. This phase of the experiment is simply a recall task.

One can also look at the final accuracy for each of the treatments, where we may be able to start to identify potential impact of the experimental variables. We can consider the difference between performance when considering the number of alternate answers shown, time pressure and the regions of problem difficulty. For example, looking at questions of high difficulty, we see that there is a sorting of performance between *alt1* and *alt3*, where the 1 alternate answer experiments show higher performance than the 3 alternate experiments. In the case of *alt3*, this is a form of selection expertise, selecting the correct answer from the four possible answers. It is expected that selection accuracy expertise in *alt1* case will be better than *alt3* because the subject is choosing between two answers rather than four answers, which inherently provides increased odds. In the *alt3* situation, the participant can learn from the feedback provided after each final response that the correct answer will also be present as one of the four answers. In the *alt1* situation, they will learn that for the majority of items the correct answer will be one of the two options, but it is not always the case. Also, the "time pressure on" results have a lower performance for tests with both numbers of alternate answers.

The number of alternatives makes a difference in final accuracy only for the difficult questions. For very difficult questions, it is likely that the initial answer is wrong and that the subject knows it is wrong, which increases the likelihood that the correct answer will be provided by the alternative and would motivate them to switch to that alternate. In *alt1*, only 50% of the time that switch will yield a correct outcome, yet final accuracy was greater than 50% for the very difficult questions. Thus, subjects had some information available to them that resulted in their ability to answer correctly more than just the random 50% by switching every time. In contrast, for the *alt3* condition, final accuracy for the most difficult questions was close to chance (25%), indicating that subjects perceived that they had little information guiding them to which of the 3

alternatives was the correct answer. By adding information, performance was reduced for these difficult items. In addition, we filtered out the cases in the *alt1* experiments where the initial accuracy was 0 and the final accuracy was 0. This case resulted in a bias in the data when comparing it with the *alt3* data.

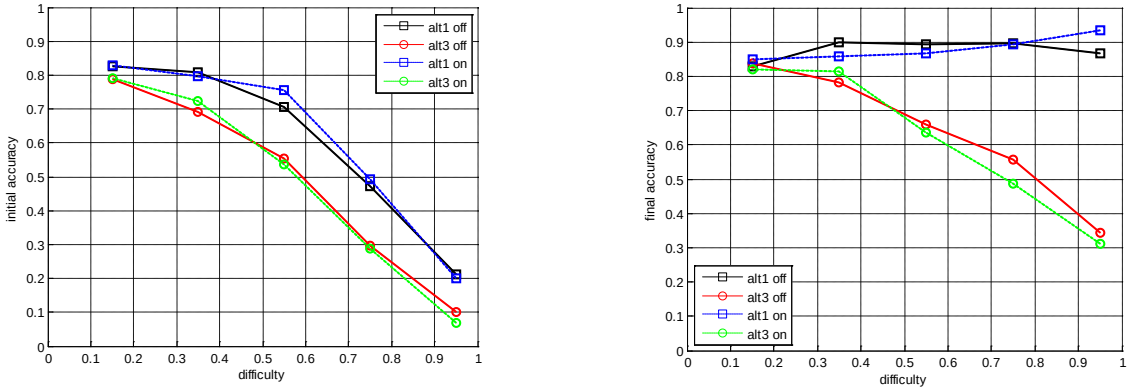


Figure 2. Initial accuracy (left panel) and final accuracy (right panel), as a function of item difficulty, for each of the four treatments.

3.2 Influence of Self-reported Domain Expertise

We now consider the impact of self-reported domain expertise of the participants (i.e., labeling a smaller number value for high expertise; 1 for highest, 2 for the medium, and 3 for the lowest) and their performance in terms of accuracy. We show the relationships between initial and final accuracy and problem difficulty.

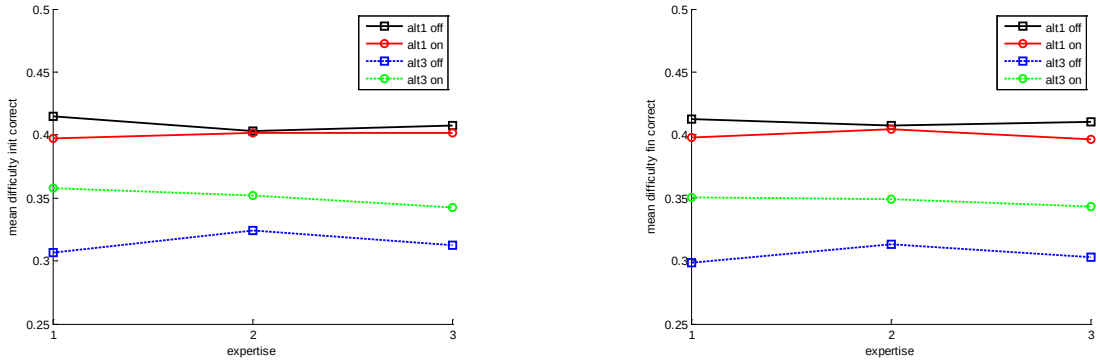


Figure 3. Four treatments (L- initial accuracy, R- final accuracy) showing average difficulty of correct initial or final responses for each treatment (high exp =1, med exp = 2, low exp = 3).

We do not see, in Figure 3, significant variation in the mean difficulty of the initial accuracy between the time pressure parameter, despite an expected increase in performance without time pressure. This is seen in the slight variation in the *alt3on* and *alt3off* cases with regard to overall difficulty. This aspect of the data requires further study. The next section contains statistical analysis of the significance of these data.

3.5. Statistical Analysis

We consider the various treatments and their impact on the accuracy of the participants. To study the statistical significance of these results, we use the Kolmogorov-Smirnov test (K-S test) [24]. This test can be used to determine if two empirical distributions are from the same distribution. These tests require a cumulative distribution function, which was obtained by considering the cumulative accuracy metrics over question difficulty in .01 increments (i.e. difficulty from 0 to 1 in steps of .01). The K-S test generates a K-S statistic, D. This is defined for functions $F_1(x)$ and $F_2(x)$, which have n_1 and n_2 sample sizes.

$$D = sup_x |F_1(x) - F_2(x)|$$

Then, significance level of the observed value of D to disprove the null hypothesis

$$\Pr(d > \text{observed}) = Q_{KS}\left(D\left(\sqrt{\frac{n_1 + n_2}{n_1 n_2}} + .12 + .11/\sqrt{\frac{n_1 + n_2}{n_1 n_2}} \right) \right)$$

So, Q_{KS} provides a measure of the probability that the distributions are the same. Below, we present the K-S statistics for several relationships.

Table 2: K-S statistics for expertise over all of the trials (top: initial accuracy, bottom: final accuracy).

	High exp	Med exp	Low exp
High exp		.0167	.0359
Med exp			.0266

	High exp	Med exp	Low exp
High exp		.0715	.0354
Med exp			.0751

We consider the distribution of the initial and final accuracy of the difficulty of for each level of self-reported expertise. In Table 2, we show the K-S statistics over all of the trials, only separated by expertise. We find that there is low significance between each of the expertise, overall. That is, there is little difference between the distributions of the difficulty of correct responses across each level of expertise.

Table 3: K-S statistics for initial vs. final accuracy for each level of expertise.

	High exp	Med exp	Low exp
Initial vs final	.5712	.6561	.2132

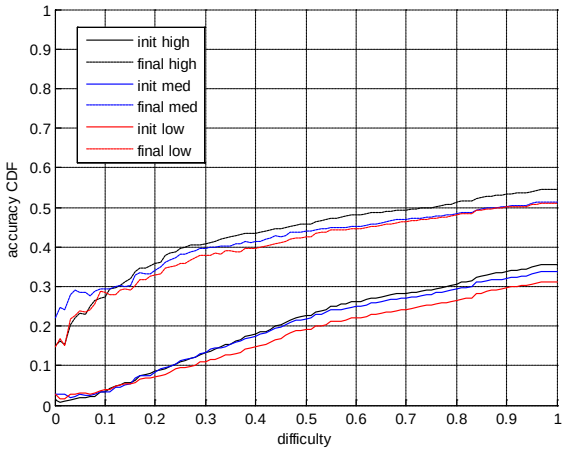


Figure 4: CDF of initial and final accuracy vs. question difficulty for the three expertise levels.

Over all of the responses, we see Figure 4 to illustrate Tables 2 and 3, where the initial and final accuracy difference is seen, but the lack of significance of the expertise parameter is shown. This plot shows the CDF of the initial and final accuracy of the responses as a function of difficulty.

Table 4: K-S statistics for high expertise over the four treatments against each treatment (top: initial accuracy, bottom: final accuracy).

	alt1off	*alt1on*	*alt3off*	*alt3on*
alt1off		.0535	.2306	.2264
alt1on			.2220	.2212
alt3off				.0309
alt3on				

	alt1off	*alt1on*	*alt3off*	*alt3on*
alt1off		.0999	.4577	.4769
alt1on			.4865	.5054
alt3off				.0546
alt3on				

In Table 4, we see the K-S statistics between the four treatments and high expertise. Here, we do not see a significant impact due to time pressure. We see slight significance between the alt1 and alt3 cases. We also see similar trends for medium and low expertise (not shown).

In Table 5, we show the K-S statistics between the initial accuracy for each of the treatments. As expected, there is statistically significant difference between the *alt1* and *alt3* cases, but not for time pressure. Table 5 (bottom) also shows the statistics for final accuracy. We would expect a larger difference/significance between *alt1* and *alt3*. We see the largest significance when time pressure is on. Last, we show in Table 7, the statistics between final and initial accuracy for each treatment. Here, the K-S statistics between the initial and final accuracy indicate a significant distinction between the two cases, as also shown in Figure 2. The difference in alt1 is greater, which may have a larger impact due to lack of confidence in their initial answer or having more confidence in the alternate answer. In the alt3 case, guessing between the 3 alternate answers if the individual has low confidence in their initial answer results in choosing between the 3 choices. This is perhaps a reason for the less significant deviation or increase in performance.

Table 5: K-S statistics for initial accuracy for the 4 treatments against each other treatment (top: initial accuracy, bottom: final accuracy).

	alt1off	*alt1on*	*alt3off*	*alt3on*
alt1off		.0287	.2034	.2048
alt1on			.2152	.2164
alt3off				.0235
alt3on				

	alt1off	*alt1on*	*alt3off*	*alt3on*
alt1off		.1581	.4612	.4733
alt1on			.5452	.5618
alt3off				.0351
alt3on				

Table 6: K-S statistics for each treatment, comparing between initial and final accuracy.

	alt1off	*alt1on*	*alt3off*	*alt3on*
Initial vs. Final	.5873	.7163	.2156	.2225

In short, we see significance in the data when comparing the *alt1* and *alt3* cases, but the difference between initial and final is not statistically significant. Additionally, as expected there is significance between the initial and final accuracy. According to these statistics, we observe that there is significance between the performance between *alt1* and *alt3*, but not as a function of time pressure and expertise. In the next section, we will adapt these findings to a model with foundations in test theory.

3.4. Expertise Mathematical Formulation

To form a model of expertise and performance into a mathematical formulation, it is necessary to come up with reasonable approximations for the data and then find mathematical representations of these relationships. First, we take the initial accuracy (same for both *alt1* and *alt3*) and the final accuracy for *alt1* and *alt3* and find linear and quadratic approximations to the data. This is shown in Figure 5.

Table 8: Regression for initial accuracy and final accuracy for *alt1* and *alt3*, expressions as a function of difficulty d and error.

Data	Regression	R
Initial accuracy	-1.2d2+.56d+.77	.034
alt1 final accuracy	.069d+.84	.01
alt3 final accuracy	-.55d2-.04d+.86	.03

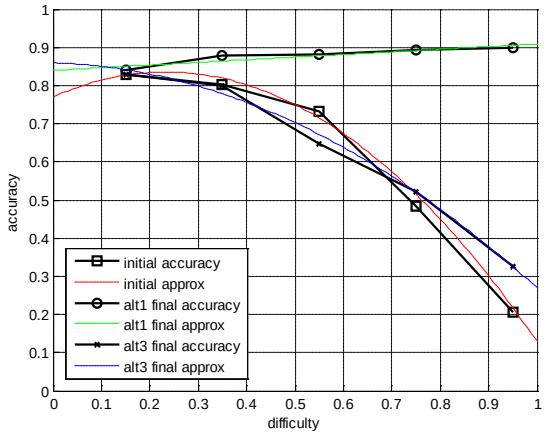


Figure 5. Initial and final accuracy for alt1 and alt3 with linear and quadratic approximations.

We borrow a concept from *item response theory* (IRT) model [25] to understand how problem difficulty can predict accuracy of responses to specific questions. The experiments from which this data came used problems with a wide range of difficulty. However, it is likely that in operational scenarios, subjects will encounter a larger variation in the difficulty of problems tested. Therefore, this model provides a means to model the likelihood of an individual being able to correctly identify the solution in various circumstances. A *three-parameter logistic (3PL) model* is used, where the probability of a correct response is determined by difficulty (b_i), discrimination (a_i), and a guessing parameter (c_i). We also modified the range r of problem difficulty. Based on this 3PL model, we choose to fit the parameters of the following expression for the probability of a correct response given subject *x*:

$$p_i(x) = c_i + \frac{1 - c_i}{1 + e^{-a_i(x - b_i)}}$$

We then take the linear approximation of the accuracy vs. difficulty curves and find the parameterization of the impulse response function (IRF) curve that fits each of these approximations. The results are shown in Figure 7 and the parameter values are in Table 6. Mapping of the experimental results to the IRF will allow for extrapolation of problems and expertise outside of what was tested in the experiments. We see a difference between the probability of correctly answering a question initially and when provided alternate answers in MSE (mean squared error). The implications of such a model are that given the difficulty of a question in an agent model, this can be used to predict the accuracy of the assessment of the agent for a particular task. Further, mapping of this task to IRF allows for comparison across other tasks in this class of questions.

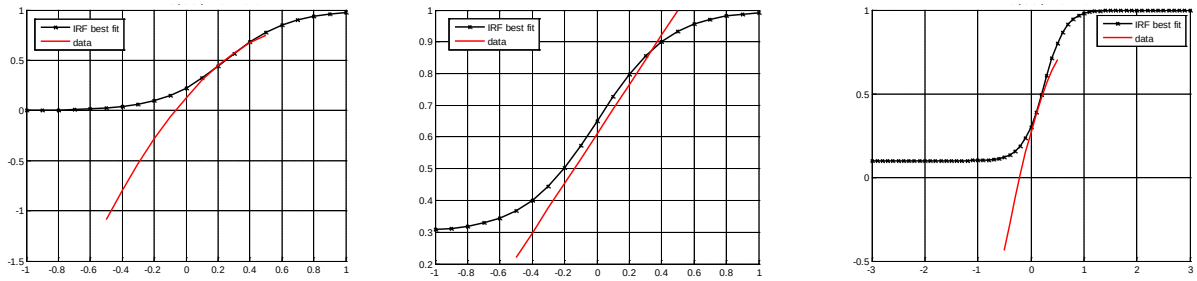


Figure 7. Initial (left) and *alt1* (middle) and *alt3* (right) final accuracy vs. problem difficulty with data and parameterized IRF best fit.

Table 9. Best fit parameters for IRF function for initial and final accuracy

	MSE	ai	bi	ci	r
Initial	3.23	5	0.25	0.0	1
alt1 Final	0.23	4.5	0	0.3	1
alt3 Final	1.21	5	0.25	0.1	3

4. Conclusion

In this work, we considered a set of experiments to study the impact of expertise on decision making ability. Characterization of this metric is important in C2 environments to assist in the design of networks to help the commanders and ultimately the organizations maximize their efficiency and decision making performance. We have characterized the probability one would expect to correctly answer a problem, given question difficulty and explored the impact of recall and selection expertise. The analysis of the data shows that there is no significant difference in performance in terms of different treatments and self reported expertise levels. The main effect appears to be problem difficulty. The only other difference in the final accuracy with 1 or 3 alternate answers, but that is hard to compare because the situations are not exactly comparable. One attempt to approach this issue is the construction of a mathematical model based on these parameters we have studied. Given data to characterize various expertise conditions, this provides us the capability to model such behavior in software agents. This serves as a beginning point to the modeling of expertise for agents in an information processing task.

References

[1] M. Ruddy, "Instantiation of a sensemaking agent for use with ELICIT experimentation," 14th International Command and Control Research and Technology Symposium, Washington, DC, June 2009,

[2] Barbara Hayes-Roth and Frederick Hayes-Roth, "A cognitive model of planning," *Cognitive Science*, vol. 3, no. 4, pp. 275–310, Oct. 1979.

[3] Sheila A. Corcoran, "Task complexity and nursing expertise as factors in decision making," *Nursing Research*, vo. 35, no. 2, March 1986.

[4] Masakazu Yagi and H. Ohno, and K. Takada, "Decision-making system for orthodontic treatment planning based on direct implementation of expertise knowledge," *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC), 2010 , pp. 2894 – 2897.

[5] M. Meyer and J. Booker, *Eliciting and Analyzing Expert Judgment: A Practical Guide*, Society of Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[6] J. M. Booker and M.A. Meyer, "Uncertainty quantification: methods and examples from probability and fuzzy theories," *2002 Proceedings of the 5th Biannual World Automation Congress*, vol. 13, pp. 135-140, 2002.

[7] David Constant, Sara Kiesler, and Lee Sproull, "What's mine is ours, or is it? A study of attitudes about information sharing," *Information Systems Research*, vol. 5, no. 4, Dec. 1994, pp. 400-421.

[8] Jennifer Chu-Carroll and Sandra Carberry, "Generating information-sharing subdialogues in expert-user consultation," *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1997.

[9] Melissa C. Thomas-Hunt, Tonya Y. Ogden, and Margaret A. Neale, "Who's really sharing? Effects of social and expert status on knowledge exchange within groups," *Management Science*, vol. 49, no. 4, April 2003, pp. 464-477.

[10] Ping Liu, Kan Liu, and Ji Liu, "Ontology-based expertise matching system within academia," International Conference on Wireless Communications, Networking and Mobile Computing, pp. 5431 – 5434, 2007.

[11] Zan Huang, Hsinchun Chen, Fei Guo, J.J. Xu, Soushan Wu, and Wun-Hwa Chen, "Visualizing the expertise space," *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, 2004.

[12] A. Mockus and J.D. Herbsleb, "Expertise Browser: a quantitative approach to identifying expertise, *Proceedings of the 24rd International Conference on Software Engineering*, pp. 503-512, 2002.

[13] D. Ma, D. Schuler, T. Zimmermann, and J. Sillito, "Expert recommendation with usage expertise," *IEEE International Conference on Software Maintenance*, pp. 535 – 538, 2009.

[14] Yanyan Li, Shaoqian Ma, Yonghe Zhang, and Ronghuai Huang, "Expertise network discovery via topic and link analysis in online communities," *IEEE 12th International Conference on Advanced Learning Technologies*, pp. 311 – 315, 2012.

[15] D. Nevo, I. Benbasat, and Yair Wand, "The knowledge demands of expertise seekers in two different contexts: knowledge allocation versus knowledge retrieval," *44th Hawaii International Conference on System Sciences*, pp. 1-10, 2011.

[16] Janet A. Sniezek and Lyn M. Van Swol, "Trust, confidence, and expertise in a judge-advisor system," *Organizational Behavior and Human Decision Processes*, vol. 84, no. 2, pp. 288-307, March 2001.

[17] Ilan Yaniv, "Receiving other people's advice: Influence and benefit," *Organizational Behavior and Human Decision Processes*, vol. 93, pp. 1-13, 2004.

[18] Janet A. Sniezek and Timothy Buckley, "Cueing and cognitive conflict in judge-advisor decision making," *Organizational Behavior and Human Decision Processes*, vol. 62, no. 2, pp. 159-174, May 1995.

[19] Siegfried Ludwig Sporer, Steven Penrod, Don Read, and Brian Cutler, "Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies," *Psychological Bulletin*, vol. 118, no. 3, pp. 315-327, Nov. 1995.

[20] Timothy J. Perfect, Emma L. Watson, and Graham F. Wagstaff, "Accuracy of confidence ratings associated with general knowledge and eyewitness memory," *Journal of Applied Psychology*, vol. 78, no. 1, pp. 144-147, Feb. 1993.

[21] Joshua L. Wiener and John C. Mowen, "Source credibility: on the independent effects of trust and expertise," *Advances* in Consumer Research, vol. 13, pp. 306-310, 1986.

[22] E. McGinnes and C. Ward, "Better liked than right: trustworthiness and expertise in credibility, Personality and Social Psychology Bulletin, vol. 6, pp. 467-472, 1980.

[23] Robert R. Harmon and Kenneth A. Coney, "The persuasive effects of source credibility in buy and lease situations," Journal of Marketing Research, vol. 19, no. 2, pp. 255-260, May 1982

[24] Numerical Recipes in C. The Art of Scientific Computing, 2nd Edition, 1992, ISBN 0-521-43108-5, Cambridge University Press, 1992.

[25] Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage Press.

U.S. Army Research, Development and Engineering Command

Providing Agility in C2 Environments Through Networked Information Processing: A Model of Expertise

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Kevin Chan, Jin-Hee Cho, US Army Research Laboratory
Sibel Adali, Rensselaer Polytechnic Institute
Jennifer Mangels, Olta Hoxha, Ksenia Lila, CUNY Baruch College
Damon Abraham, CUNY Graduate Center

ICCRTS 2014
Track # 4
Paper # 33
June 17, 2014

Problem

- Decision making in command and control (C2) environments requires gathering, processing and sharing of all available information to establish intelligence and to increase situation awareness.
- Complex criteria play a role in such a decision:
  - trustworthiness (reliability, benevolence) of the sources
  - ability (expertise) of analysts to analyze and filter the information
  - underlying difficulty/complexity of the decision space
  - other network level factors like social influence

US ARMY RDECOM

Contribution

ARL

- Summarized experiments by CUNY that study the expertise of individuals who answer general knowledge domain questions in an experimental setting
- Studied the expertise and difficulty component of this process
- Incorporated an initial mathematical model of expertise and difficulty into an agent model for information processing in networks

Connections to C2 and ELICIT

Command and Control (C2)

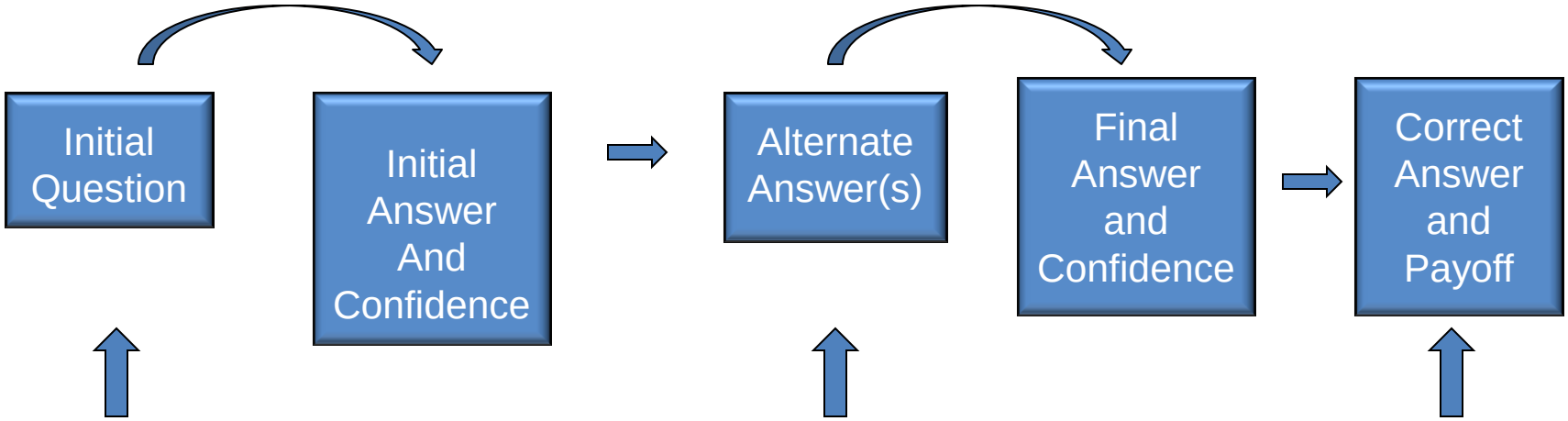
- Developed distributed information sharing framework that enhances situation awareness
- Predicted correctness of information sharing, given specific level of problem difficulty

Agent Models

- Developed ELICIT- based and other agent based models for information sharing scenario
- Predicted correctness of information processing, given various levels of problem difficulty

Experiment Flow

- Set of 120 questions, titrated to achieve initial accuracy of .33



Test	Treatment	Abbreviation	# of subjects
(1-a)	One alternate answer with no time pressure	*alt1off*	33
(1-b)	One alternate answer with time pressure	*alt1on*	39
(3-a)	Three alternate answers with no time pressure	*alt3off*	39
(3-b)	Three alternate answers with time pressure	*alt3on*	37

Table 1: Description of experiment treatments and number of subjects tested.

- Cleaning of data, titration breakers, confused participants

Experiment Setup

Demographics

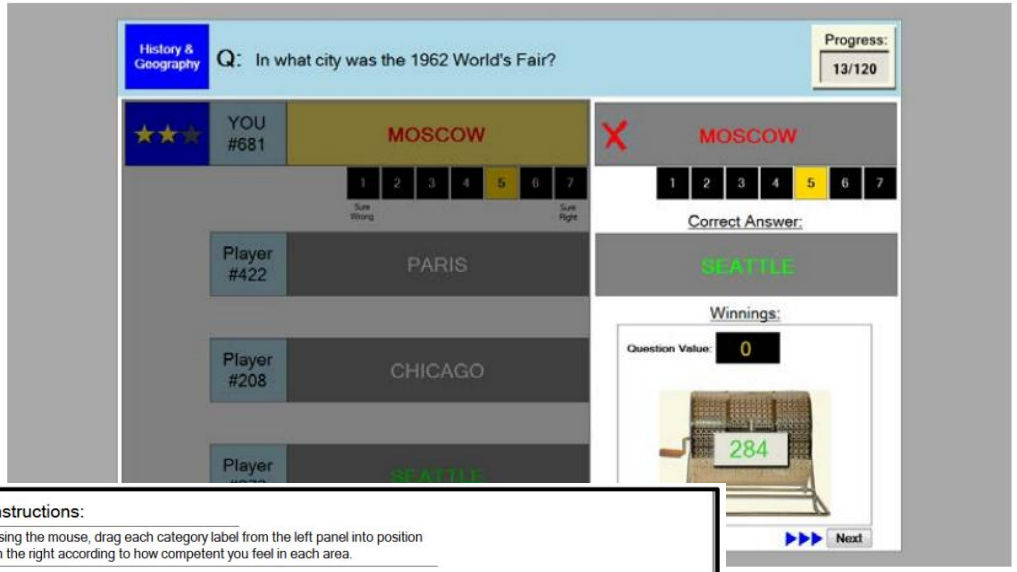
- College students
- Expertise of domains

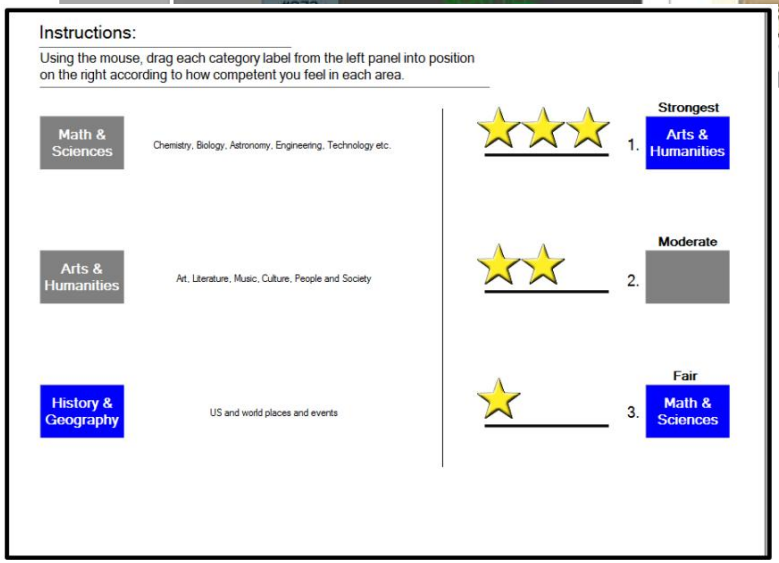
Metrics

- Initial and Final Accuracy
- Initial and Final Confidence

Treatments

- Alternate answers: {1, 3}
- Time pressure: {on, off}

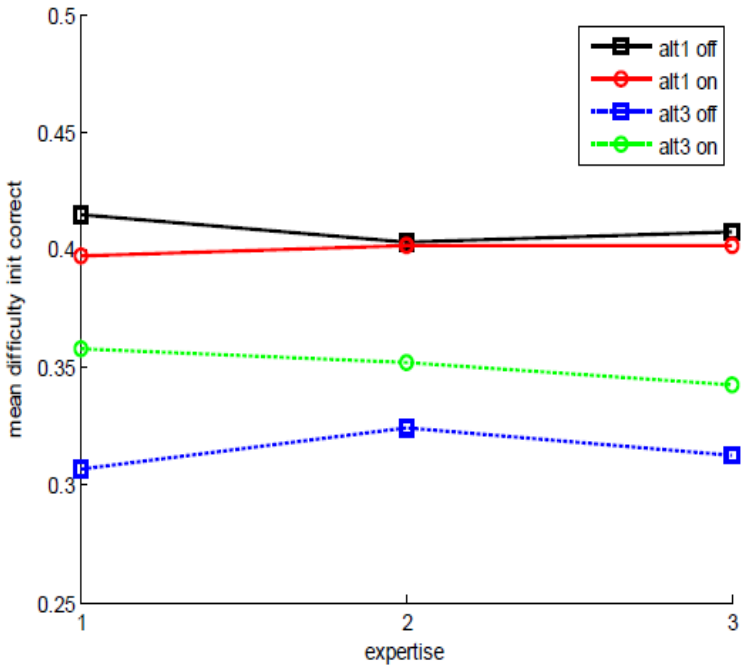


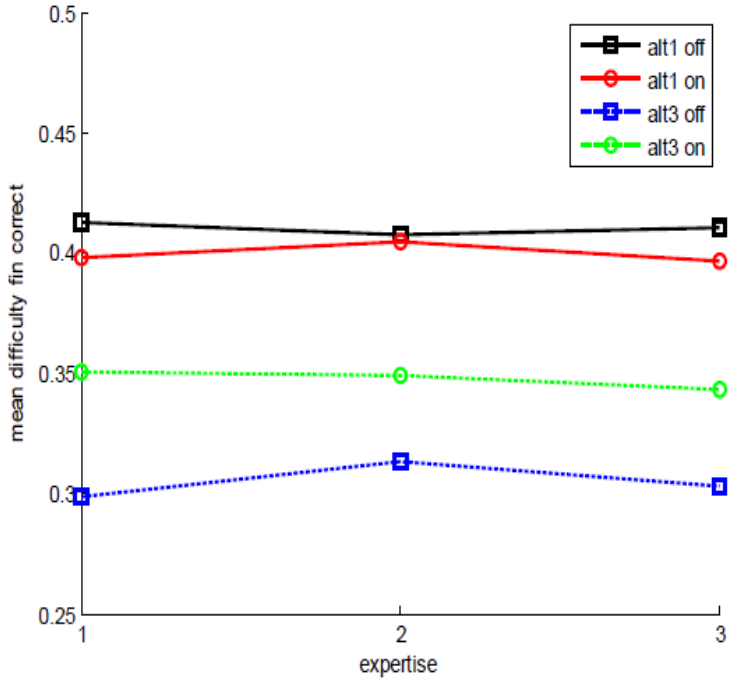


TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Correctness vs. Expertise

- Mean difficulty of correct accuracy vs. expertise for each treatment





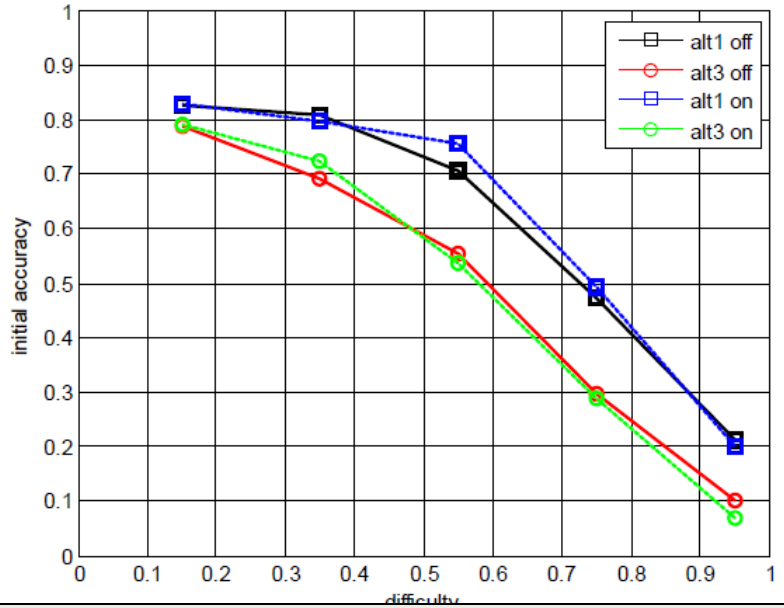
Correct answers don't exhibit significant differences when it comes to self-reported expertise

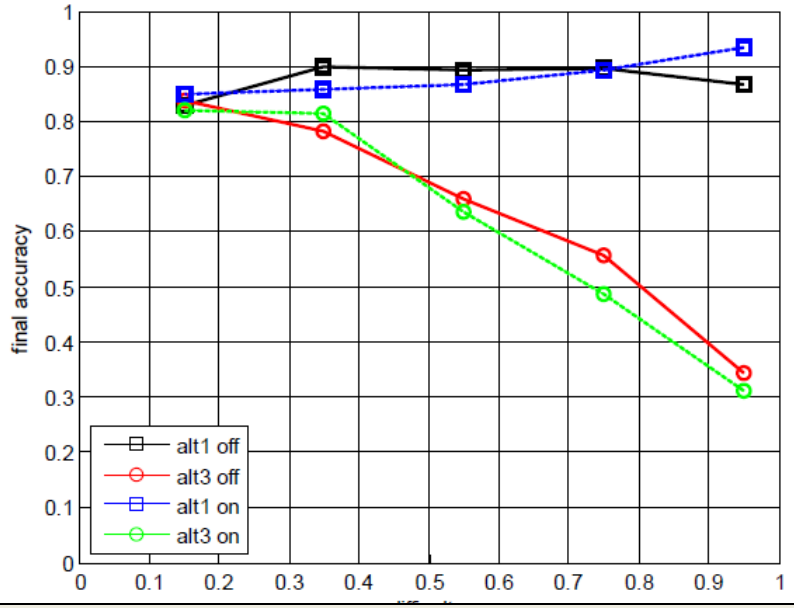
US ARMY RDECOM

Accuracy vs. Problem Difficulty

ARL

- Initial and Final accuracy versus problem difficulty
- Data Cleaning: Initial accuracy (removed wrong|wrong for alt1)





Based on the number of alternate choices vs. problem difficulty, we see an increase in accuracy

Statistical Analysis

- Kolmogorov-Smirnov test (K-S test) determines if two empirical distributions are from the same distribution
- CDFs, $F_1(x)$ and $F_2(x)$

$$D = sup_x |F_1(x) - F_2(x)|$$

- Q_{KS} is a measure of the probability that the distributions are the same

$$\Pr(d > \text{observed}) = Q_{KS}\left(D\left(\sqrt{\frac{n_1 + n_2}{n_1 n_2}} + .12 + .11/\sqrt{\frac{n_1 + n_2}{n_1 n_2}} \right) \right)$$

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Statistics between Treatments

Expected (lack of) significance for initial accuracy, difference in final accuracy

- Initial Accuracy (less significant difference in performance)

	alt1off	*alt1on*	*alt3off*	*alt3on*
alt1off		.0535	.2306	.2264
alt1on			.2220	.2212
alt3off				.0309
alt3on				

- Final Accuracy (significant difference in performance)

	alt1off	*alt1on*	*alt3off*	*alt3on*
alt1off		.0999	.4577	.4769
alt1on			.4865	.5054
alt3off				.0546
alt3on				

Statistics between Expertise

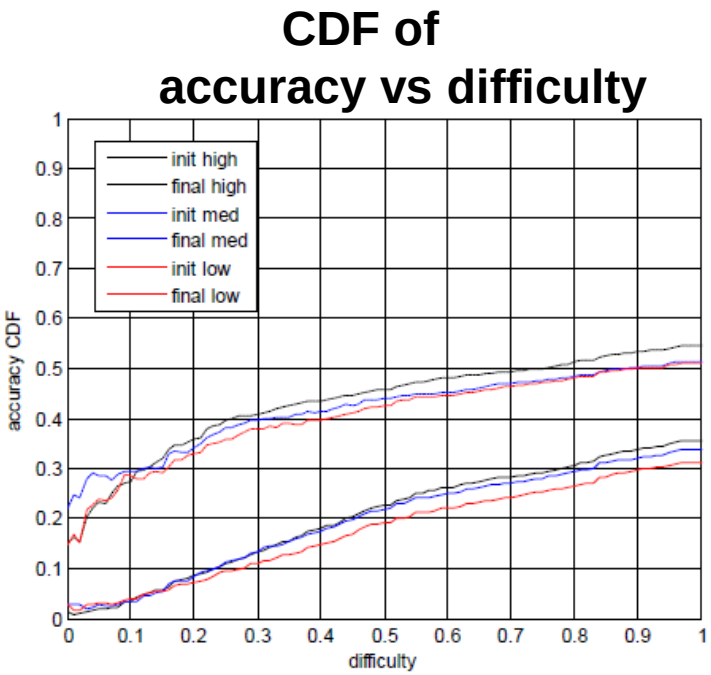
Minor statistical significance between accuracy for different expertise

- Initial Accuracy (little statistical significance)

	High exp	Med exp	Low exp
High exp		.0167	.0359
Med exp			.0266

- Final Accuracy (little statistical significance)

	High exp	Med exp	Low exp
High exp		.0715	.0354
Med exp			.0751



TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Experiment Conclusions

- Significance is observed:
  - when comparing the *alt1* and *alt3* cases
  - between the initial and final accuracy

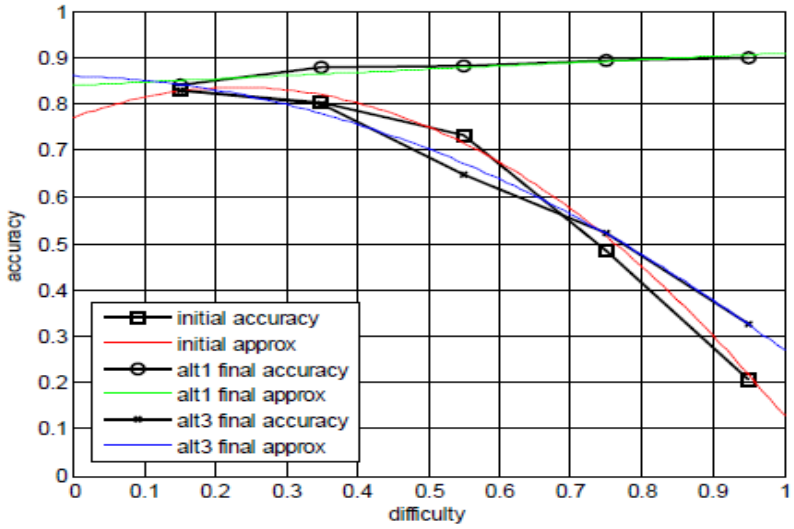
- No significance is observed with respect to the function of time pressure and expertise

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Mathematical Formulation

- Regression of initial and final accuracy can inform agent models of accuracy given problem difficulty

Data	Regression	R
Initial accuracy	$-1.2d^2 + .56d + .77$	.034
alt1 final accuracy	$.069d + .84$	.01
alt3 final accuracy	$-.55d^2 - .04d + .86$	



TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Item Response Theory (IRT)

- Use IRT to understand how problem difficulty can predict accuracy of responses to specific questions

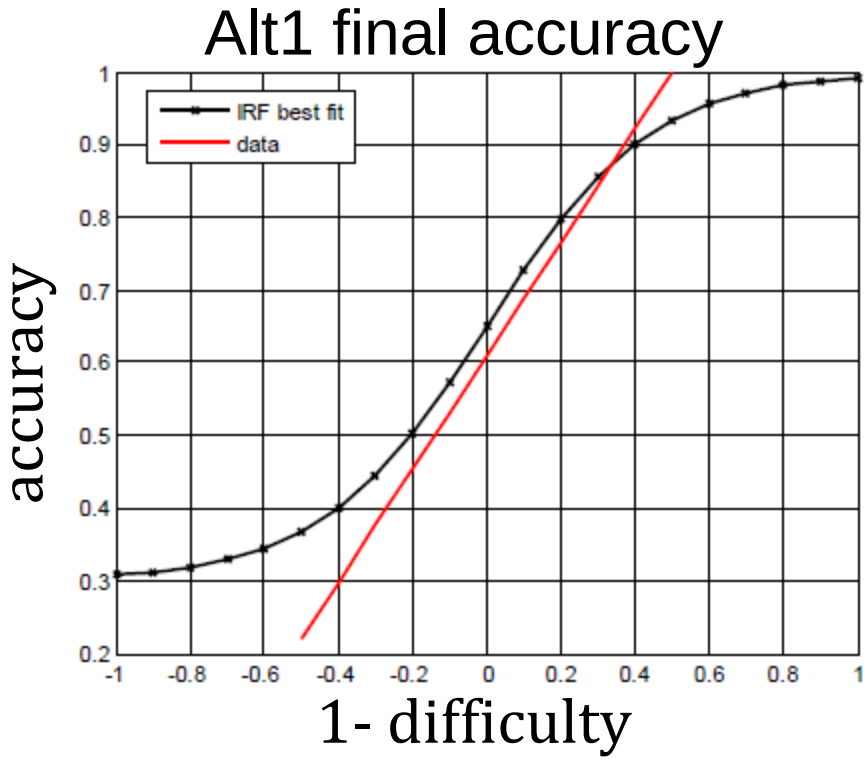
- Three-parameter logistic (*3PL*) model

$$p_i(x) = c_i + \frac{1 - c_i}{1 + e^{-a_i(x - b_i)}}$$

b – difficulty

a – discrimination, slope

c – pseudo-guessing



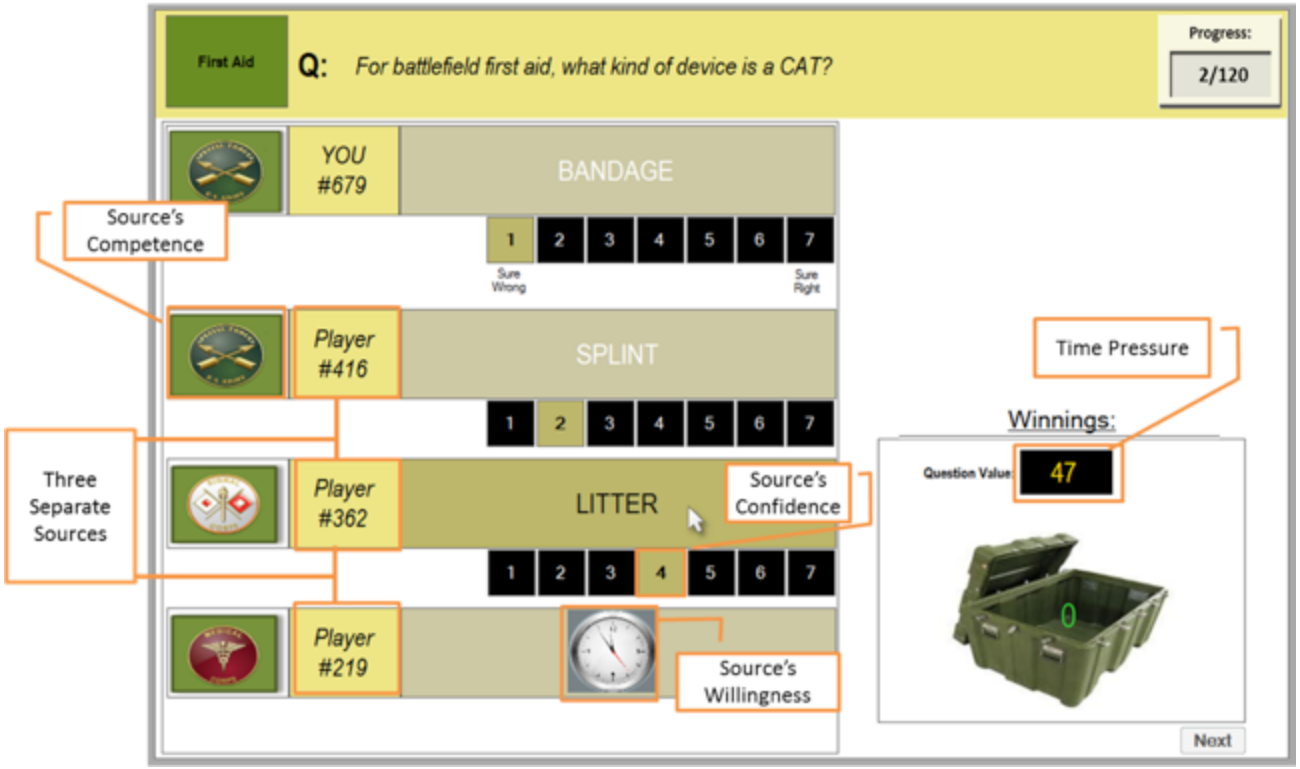
Conclusions

- Expertise / training impacts decision making ability; however, observations do not show any statistical significance in decision making performance
- Problem difficulty vs. decision making performance shows statistical significance
- Mathematical model based on experimental parameters provides capability to configure software agents to match the trends observed in the human experiments

UNCLASSIFIED / APPROVED FOR PUBLIC RELEASE

Military Relevant Experiments

- Create a military relevant study
- Compare against existing normative study



TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Thank You!

Contact us at:

Kevin Chan

U.S. Army Research Laboratory

kevin.s.chan.civ@mail.mil